AFRL-IF-RS-TR-2002-124
Final Technical Report
June 2002

GENERIC EVENT RECOGNITION AND EXTRACTION (GERE)

Syracuse University

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. L142

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2002-124 has been reviewed and is approved for publication.

APPROVED: Raymond A. Liuzzi

RAYMOND A. LIUZZI
Project Engineer

FOR THE DIRECTOR:

MICHAEL L. TALBERT, Technical Advisor
Information Technology Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
JUNE 2002

3. REPORT TYPE AND DATES COVERED
Final Mar 01 – Dec 01

4. TITLE AND SUBTITLE
GENERIC EVENT RECOGNITION AND EXTRACTION (GERE)

5. FUNDING NUMBERS
C - F30602-01-2-0517
PE - 62301E
PR - EELD
TA - 00
WU - 01

6. AUTHOR(S)
Elizabeth D. Liddy and Eileen E. Allen

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Syracuse University
Center for Natural Language Processing
School of Information Studies
4-206 Center for Science and Technology
Syracuse New York 13244

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington Virginia 22203-1714

AFRL/IFTD
26 Electronic Parkway
Rome New York 13441-4514

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2002-124

11. SUPPLEMENTARY NOTES
AFRL Project Engineer: Raymond A. Liuzzi/IFTD/(315) 330-3577/ Raymond.Liuzzi@rl.af.mil

12a. DISTRIBUTION / AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

Generic extraction was the primary focus of this funded project. To successfully extract relationships about generic entities and events, the previously developed Named Entity-based extraction algorithm was modified and extended into a process to work on events and nominal entities within the Nuclear Smuggling data domain. A secondary focus was the identification and categorization of numeric and time expressions. An evaluation shows significant improvement alone precision and recall: measures. The framework and a preliminary algorithm for specialization of event frames and attributes was designed.

14. SUBJECT TERMS
Information Extraction, Natural Language Processing, EELD

15. NUMBER OF PAGES
11

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

Generic Event Recognition and Extraction

Background
During the initial EELD Seedling Project, the Center extended its NLP-based entity, concept, and relation identification and extraction capabilities to include events – based on a modeling of the two expressed domains of interest, i.e. mergers & acquisitions, and smuggling. In developing the event extraction capability, a generic model of 'change & effect' was used, which was then specified by studying a sample of reports first for mergers & acquisitions, and then nuclear smuggling. These specialized models enabled identification of the important events and sub-events for which the extraction system needed rules. However, it was found that both the rule development process and the system's extraction module could be improved in effectiveness and efficiency if a more generic event recognition and extraction capability was first developed, and then used as the basis for specialized event extraction.

The following report outlines the work accomplished through the project towards that goal.

Testbed
Five document collections were downloaded from the DARPA/ SHIELD web site which deal with nuclear smuggling. These collections were prepared for processing, and testing. A total of 4888 documents were collected. About 10%, or 503 documents, were set aside as our test set. The remainder were used for training and development.

Baseline Evaluation
A baseline evaluation of performance was done on these documents using the processing system before development of a robust set of generic extraction rules. Elements that were evaluated include the identification and categorization of Proper Named Entities and Numeric Concepts, and the extraction of events, entities and relations among them.

Evaluation was based on the standard information metrics of precision and recall, that is, accuracy and coverage of the system. The results of the baseline evaluation were later compared to the Final Evaluation results. Detailed results from both evaluations are discussed under Final Evaluation.

EELD Coordination
Throughout this development period, it was important to work closely with other

participants in DARPA's EELD program to develop a database schema to facilitate interaction among the various technologies. We responded in an effort to ensure that the schema remained flexible to allow for unknown data and unexpected forms that are discovered through the extraction process to be incorporated and thus usable by Link Discovery and Pattern Recognition technologies.

System Development

Numeric Expressions

The correct identification and categorization of names and numeric concepts is important for the process of extracting relationships from text. The smuggling data includes named entity types new to our system, and refinements in both identification and categorization were warranted.

Numeric concept identification and categorization was a newer process. Based on the results of the baseline study, development of rules to improve numeric concept detection and categorization was among the first tasks addressed, in particular, to improve the detection of dates and time periods which can be quite complex. This was done by adding new rules and knowledge sources to the generic entity-based system.

Generic Event Extraction

To successfully extract relationships about generic events and entities, the extraction algorithm was modified and extended into a process to work for all of our previous Named entity-based extractions as well as the generic event and entity extractions.

Rules for generic event extraction were written using an iterative process of development and testing. The initial focus of the rule set was on the direct extraction of agency and object relationships based on verb occurrences, that is, who or what initiated an event and who or what was affected by the event, with particular attention to the problem of active and passive voice. Previously, some of this information was extracted through the Named entity-based extraction algorithm, but without the flexibility and power that is now available. A second focus was on the development of rules which recognize when nominalization forms of events occur, such as "explosion" as an instance of "explode", or "assassination" as an instance of "assassinate".

Additional extractions which provided attributes to the events, such as location, point in time, and other links were developed.

Specialization

The predominant events within the collection have been analyzed for specialization of events. A preliminary outline of event classes of interest was developed (Appendix A). For these event classes, four were selected to test possible algorithms for specialization: arrest, kill, smuggle and detain. For this purpose, possible use of FrameNet was pursued. However, FrameNet was not intended for the type of data EELD is interested in, and frame representations were not available for the events of interest. We therefore proceeded to develop specialized slot values for the four test event classes.

The following provides an example of the result of event class identification and specialization.

Generic Table:

```
Frametype = event
Text = assassination
        Agent = criminal group
        Object = Viktor Novosyolov
        Point in time =  Oct. 20, 1999
        Location = St. Petersburg
```

Specialization:

```
Frametype = event
Text = assassination
Event type = kill
        perpetrator = criminal group
        victim = Viktor Novosyolov
        Point in time =  Oct. 20, 1999
        Location = St. Petersburg
```

An algorithm has been developed which utilizes rules to specialize the four events and will provide the foundation for further development and testing within the larger EELD program.

Final Evaluation

A final evaluation was carried out in order to report improvement statistics. Since the evaluation itself is a manual process, it was impractical to evaluate the 503 documents within the test set, so two subsets of documents were chosen. The evaluation was based on 26 documents which had a total of 299 sentences. Two-thirds of the documents had been used for the baseline evaluation. This provided a review of the improvements made on data that had been seen, even though it was not used for development. The remainder were new test documents, and represented different and a wider variety of data sources. Evaluation of this group demonstrated the performance of the system as applied to new unseen data, and can suggest performance as new domains of interest are tackled. There was no significant performance difference between the two groups of documents.

A complete evaluation was undertaken, using the measures of precision and recall to determine the performance of the Named Entity and Numeric Concept bracketing and categorization processes as well as the performance of relation extraction.

Precision and recall are standard measures of performance within the information field. Precision is the relative accuracy of the system, and recall is the relative coverage of the system. Both are expressed as a percentage. They are defined as follows:

$$\textit{Precision} = \frac{\textit{number of correctly extracted items}}{\textit{total number of linked items}}$$

$$\textit{Recall} = \frac{\textit{number of correctly extracted items}}{\textit{total number of items in the sample}}$$

Results:

Bracketing – Named Entities and Numeric Concepts

It can be seen that the largest improvement was for the identification of numeric concepts, which reflects a major focus of this project.

	Baseline	Final	Improvement
Named Entities			
Precision	92%	92%	0%
Recall	92%	94%	2%
Numeric Concepts			
Precision	81%	93%	15%
Recall	82%	88%	7%

Categorization – Named Entities and Numeric Concepts

Although bracketing of names appears to have improved somewhat and categorization of names appears to have degraded somewhat, the small percentage may be an artifact of the particular sample that was selected since it was not a large sample. However, the small increase in categorization of numeric concepts is significant since a significantly larger proportion of numeric concepts were available for categorization as a result of the improvement in bracketing.

	Baseline	Final	Improvement

Named Entities			
Precision	97%	97%	0%
Recall	77%	75%	-3%
Numeric Concepts			
Precision	90%	94%	4%
Recall	90%	94%	4%

Extraction of Entities and Events

All entities (expressed either as a Proper Named Entity or as a common noun or noun phrase) and events were evaluated for extraction performance if an attribute was associated with them. For example, in the phrase " the cat ran to the barn", cat would not be an extractable entity because there is no attribute related to the cat. However, in the phrase "the calico cat ran to the barn", "cat" is extractable, with the attribute "characteristic = calico". Both phrases have the extractable event "run", with an "agent" slot and a "destination" slot.

Recall of entities and events was measured to indicate the proportion of extractable events and entities that were extracted. While this measure says nothing about the quality of attributes that were extracted, it is clear that much more information was extracted.

Event and Entity extraction (Recall):

	Baseline	Final	Improvement
Events:	50%	93%	86%
Entities:	27%	61%	126%

Attribute performance was measured in three ways: as related to entities; as related to events; and as related to both.

Relation/Attribute extraction:

	Baseline	Final	Improvement
Entity Attributes:			
Precision	71%	87%	23%
Recall	13%	44%	238%
Event Attributes:			
Precision:	64%	69%	8%
Recall:	22%	48%	118%
All Attributes:			
Precision	68%	75%	10%
Recall	20%	46%	130%

What is noteworthy about these results is that both precision and recall figures improved significantly for extractions. It is well known within the field of information retrieval, for example, that it is typical that as precision of results increases, recall decreases. However, in the information extraction arena, it is clear that improvements can be seen in both coverage and in accuracy.

Future Work

There are clear mandates for future work as a result of this evaluation. Future projects, in particular the big EELD project, will require a continued effort to increase recall of extractions as well as a larger specialization effort. In addition, efforts to reduce the error in the resultant extractions to boost precision (accuracy) are warranted.

Appendix A: Potential Events for Round 1 Specialization & their variants

Acquire
Acquire
Acquisition
Capture
Confiscate
Harvest
Nab
Obtain
Obtained
Own
Possession
Possess
Blackmail
Procure
Procurement
Purchase
Receive
Regain
Purloin
Seize
Seizure
Siphon
Siphon_off
steal
Theft
Trade
Trade_in
take_over
Hand_over
Offer

Affiliation
Join
Affiliate
Belong_to

Kill
Assassinate
Execute
Murder
Kill
Terminate

Accuse
Accuse
Charge
Cite
Condemn
Indict

Arrest
Apprehend
Arrest
Convict
Detain
Detention
Halt
Jail
Lock_up
Shut
Shut_down
Stop

Transport
Carry
Carry_out
Channel
Deliver
Delivery
Pack
Package
Pass
Pass_through
Smuggle
Smuggling
Ship
Shipment
Transfer
Transmit
Transport
Transportation
Travel
Travel_to
Visit
Walk

Support
Aid
Arm
Assist
Assistance
Cooperate
Coordinate
endorse
Equip
Establish
Facilitate
Fund
Funding
Funds
Help
Hire
Invest
Investment
Rebuild
Permit
Pledge
Praise
Protect
Protection
Provide
Provide_for
Provision
Provisions
Reward
Rescue
Reinforce
Restore
Restoration
sponsor
Strengthen
Supplier
Supply
Support
Supporter
Sustain
Take care

Meet
Link
Negotiate
Meet
Meeting

Attack
ambush
Assault
Attack
Bomb
Bombing
Break
Break_in
Break_into
Burn
Cut_off
Defeat
Destroy
Destruction
Detonate
Detonation
Eliminate
Explosion
Fight
Fighting
Fire_at
Get_rid_of
Hit
Infect
Inflict
Kick
Launch
Lay_waste_to
Poison
Provoke
Shell
Ransack
Strike
Target